Volume 13 Number 8 1985 Nucleic Acids Research

Analysis of an inversion within the human beta globin gene cluster

M.W.Jennings, R.W.Jones, W.G.Wood and D.J.Weatherall

MRC Molecular Haematology Unit, Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK

Received 14 March 1985; Accepted 22 March 1985

ABSTRACT

We have cloned and sequenced the DNA from two regions of the defective β-globin gene cluster from a patient with Indian $A\gamma\delta\beta$ thalassaemia, and confirmed the complex and unusual pattern of rearrangement involving two seperate deletions (0.8 kb and 7.5 kb) the inversion of the 15.5 kb segment seperating them, as previously proposed from gene mapping studies [1]. All four breakpoints occur within the transcribed region of the globin genes and at one junction are found six nucleotides of unknown origin. This unique rearrangement results in enhanced expression of the upstream fetal gene, and is therefore is pertinent to the localisation of any putative control region involved in the coordinate expression of fetal and adult genes.

INTRODUCTION

There are a number of deletions within the human β-globin gene cluster which bring about δβ-thalassaemia and related disorders, resulting in the persistent expression into adult life of the upstream fetal γ genes. Most are simple linear deletions involving the δ and β, and sometimes the $A\gamma$, genes [2]. However one particularly interesting rearrangement has been described in Indians which, on the basis of gene mapping, appears to involve two deletions flanking an inverted segment of this cluster [1]. A natural rearrangement involving a localised inversion has not previously been demonstrated in eukaryotic DNA so the analysis of this case is of interest both in terms of mechanism by which it arose and its effect on the regulation of globin gene expression.

Amongst the various deletions within the cluster there is much interest in trying to relate the persistence and degree of γ gene expression to the pattern of rearrangement or loss of particular sequences, and thereby identify any regions which may play a regulatory role in the normal expression of fetal and adult genes. The analysis of this inversion-deletion rearrangement shows that it is unlike any of the other deletions described in the β complex, removing two segments, and inverting intact the entire region between the δ and $A\gamma$ genes. This therefore further constrains the various models proposed to rationalise the molecular basis for enhanced γ gene expression.

MATERIALS AND METHODS

DNA was prepared from the peripheral blood leukocytes of an Indian patient (BK) previously shown to be homozygous for $(A_\gamma\delta\beta)^o$ thalassaemia [3] and who was the subject of restriction enzyme mapping by Jones et. al. [1]

Southern blot analysis and molecular cloning of this DNA was carried out using methods described by Maniatis [4]. The DNA was digested to completion with EcoRI and size fractionated by centrifugation through a sucrose gradient. Fragments of 3 to 5 kb were ligated into the EcoRI sites of Charon 3A vector "arms" which had previously been treated with alkaline phosphatase. These recombinants were packaged into phage particles and the resulting library screened seperately with an 866 nucleotide EcoRI/PstI fragment probe from the 3' end of the β gene, and a 2.3 kb PstI fragment probe containing the δ gene, both cloned from genomic DNA of a normal individual. Inserts identified by these probes were subcloned into the EcoRI site of pBR322 and analysed by restriction mapping.

The nucleotide sequence was determined according the method of Maxam and Gilbert [5], and analysis of the sequence results in relation to sequences previously published for the A_γ, δ, and β genes [6,7] were made with the assistance of computer programmes developed by Staden [8] and data from the EMBL nucleotide sequence data library .

RESULTS AND DISCUSSION

Cloning and nucleotide sequence analysis of the breakpoints.

EcoRI genomic DNA digests from BK produce two abnormal fragments of about 4.2 kb which are distinguishable on Southern blots by their ability to hybridise to either δ or β gene-specific probes ,and which appear to span the ends of the inverted segment [1]. A recombinant Charon 3A bacteriophage library was generated using 3-5 kb EcoR1 fragments from BK and screened with seperate δ and β globin gene specific probes. Inserts of approximately 4.2 kb were isolated, subcloned into pBR322, and designated pBKδ and pBKβ according to the probe with which they hybridised. Results of restriction mapping of these plasmids were consistent with each of the EcoR1 fragments postulated to contain the ends of the inverted sequence (figure 1).

Fragments from these plasmids containing the junctions of the rearrangement were identified from either genomic mapping data or by comparison of fine structure maps of the BK clones with the corresponding regions cloned from unrearranged β-clusters, and the nucleotide sequence determined for both strands across the junctions.

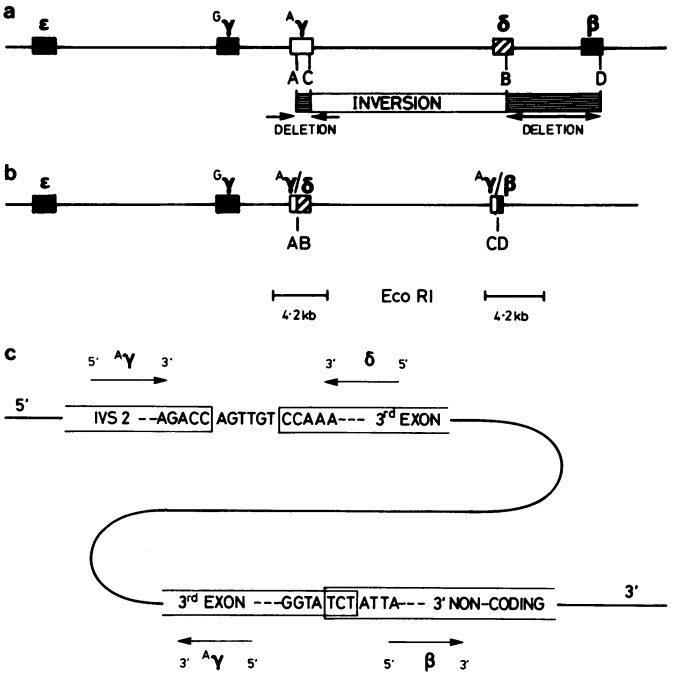


Figure 1.
The β globin gene cluster showing the breakpoints A,B,C, and D (a) which are rearranged in patient BK (b) as an inversion of the 15.5 kb segment between C and B and deletions of 834 bp of the $^{A}\gamma$ gene between A and C, and of 7.46 kb between sites B and D in the δ and β genes. Thus, sequences from the 5' end of the δ gene at C are found attached to the $^{A}\gamma$ gene sequence at A and sequences from the 3' end of the $^{A}\gamma$ gene at B found attached to the 3' end of the β gene at D (c). The two cloned EcoR1 fragments are also indicated (b).

The nucleotide sequences found at the junctions are shown in figure 2 together with the corresponding sequences of the $^{A}\gamma$, δ and β genes from which they are derived. In pBKδ, the coding strand of the $^{A}\gamma$ gene remains intact up to nucleotide 151 in IVS2, where it is followed by six nucleotides (AGTTGT) of unknown origin before continuing with the inverted sequence of the non-coding strand of the δ gene, starting at nucleotide 44 in the third exon and working towards what would normally be the 5' end of this gene. The junction region in pBKβ consists of the inverted sequence of the $^{A}\gamma$ gene, from its 3' end up to nucleotide 116 of the third exon, fused to the 3' non-coding region of the β gene 23 nucleotides beyond the termination codon.

These results therefore confirm the original interpretation of the gene mapping of this rearrangement, produced by two deletions, one of 834 nucleotides in the $^{A}\gamma$ globin gene and one of 7460 nucleotides in the region

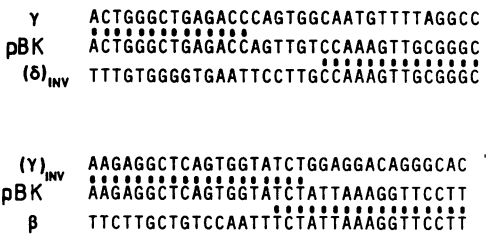


Figure 2. The nucleotide sequences determined at the junctions of the rearrangement (centre line of triplets, pBK) are shown with the normally contiguous parental sequences (above and below BK) which have been deleted in BK. Sequences published for the $A\gamma$ gene show a polymorphism (C or T) 7 bp 3' to the breakpoint A [34]. Sequences similar to the consensus immunoglobulin signal (CACAGTG) occurr at breakpoint A (CCCAGTG), and slightly to the left of C (CTCAGTG) There is patchy homology between sequences around A and B where, by staggering the breakpoints by 6 bp and allowing two single base loops in sequence A, there are 22 homologies in a stretch of 33 nucleotides. There is little homology between sequences at C and D other than TCT shared at the breakpoints, although there is a 12 bp perfect homology approximately 240 bp 5' to the breakpoints. Sequences around C and D do individually show some homology with the other junction; the 13 bases immediately to the left of the breakpoint D and at the AB junction share 11 homologous bases, and 19 bases from the left of the breakpoint B and to the left of the AB junction share 14 homologous bases, including the sequence AGTGGT near B which is similar to the 'orphan' AGTTGT. Two direct repeats of 4 and 8 bp were found in the sequence around B, the former forming part of the sequence AGTTGC with the repeat TTGC at the breakpoint, which again may be related to the bases found at the junction.

between the δ and β genes; the 15.5 kb between these two deletions remains present, but has been inverted.

The 'orphan' hexanucleotide.

The origin of the sequence AGTTGT, found at the left hand junction, is unclear as this sequence is not found at the corresponding positions in the parental sequences. Short stretches of related sequence do occur in the adjacent parental sequences from which these nucleotides could be derived, by a mechanism similar to that proposed for the similar novel heptanucleotides found in the Indian β^o-Thalassaemia [9], but the circumstantial evidence for this is very weak. In any case such orphan nucleotides are not uncommon, being found at the sites of recombination of immunoglubulin genes, other deletions in the globin complexes, and recombination of virus sequences in mammalian cells [10-13]. It is hard to assign any origin to the short sequences of 5 or 6 nucleotides, but in some cases where longer stretches of DNA have been inserted at the recombination breakpoints, they appear to be fragments of the deleted segment [11,12]. It would seem therefore that these nucleotides may be incorporated during the repair of the breakage from oligonucleotides generated during the event, or possibly by some template-independent terminal transferase activity.

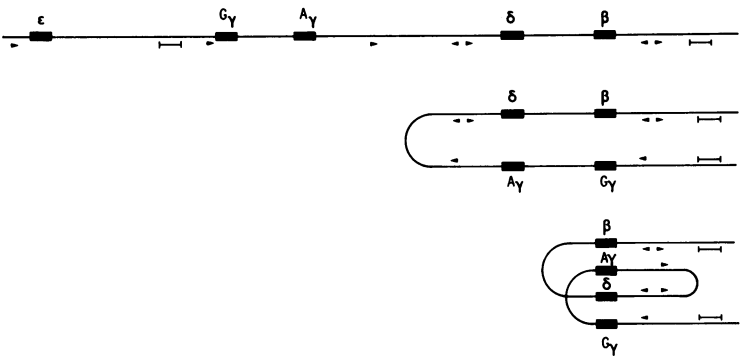


Figure 3. Folding of the beta globin cluster. The cluster is doubled back on itself, forming a hairpin with the inter-A_γ-δ region at the hinge, such that the Alu repeats adjacent to the G_γ and β genes, and A_γ and δ genes, and the two KpnI family repeats which form an inverted pair, are aligned . The corresponding genes then also lie adjacent, though in opposite orientations. If the loop is folded once more such that the Alu repeats adjacent to the δ and A_γ genes are aligned with the Alu sequences adjacent to the G_γ and β genes, all four globin genes lie adjacent. In this alignment the breakpoint of the BK rearrangement in sequcence A lies within 230 bp of the breakpoint in B and that breakpoints in C and D are within 700 bp.

Origin of the rearrangement.

The complex nature of this rearrangement raises questions about its origin; whether any features of the nucleotide sequences involved may have influenced the pattern and precise positions of the recombination, and what might be responsible to bring about the overall folding of the cluster required to bring the recombining sequences together.

The formation of a loop within the cluster, stabilised by Alu repeat sequences could bring the A_γ gene into close proximity with both the δ and β genes as originally proposed [1]. The identification and precise postioning of further Alu and KpnI family repeats [6,7] suggests an alternative folding model to bring the breakpoints together, which discloses an underlying symmetry in the distribution of repetitive sequences and globin genes within the cluster, see figure 3 (R.Nicholls, in preparation).

A number of features have been suggested as being characteristic of illigitimate recombination events, such as short regions of identity between the recombining sequences, but none have proved to be consistent, and the low number of independent recombinations within eukaryotic genomes thus far characterised (other than immunoglobulin genes) leaves this topic very speculative [14-16]. In common with two other deletions in the globin gene clusters and in other systems [9,11,12,17,18], there is some patchy homology around the breakpoints at the 5' junction (A_γ-δ_{inv}), but not between the

sequences involved in the 3' junction ($^{A}\gamma_{inv}$-β). The 3' junction does show some characteristics found in an example of an inverted aposition of histone sequences [19], namely a small overlap region and the presence of short direct repeats. Some of the findings of the computer comparisons are outlined in the legend to figure 2, but overall they show no consistent features.

It seems noteworthy that all four breakpoints have occurred within genes, three near to the termination codon, which represent only a small proportion of the region, and that they lie similar distances from adjacent Alu repeat sequences. A relationship between breakpoints and gene coding regions has also been noted in a rearrangement of histone genes in which parts of the normal repeat structures are found in inverted apposition,probably due to a larger inversion; the two breakpoints lie between the 'TATAA' and 'CAAT' boxes 5' to the coding region [19]. In another example of histone gene rearrangement, producing an inverted duplication, the boundaries coincide with conserved promoter elements [20]. These observations suggest two interpretations; the nucleotide sequence homology may permit alignment of the genes and the subsequent rearrangement having the unusual outcome of producing an inversion, and secondly, that the interaction of the sequences is somehow dependent upon physical characteristics of regions with the potential for transcription.

Recombination with inversion seems to be a rarity, at least within eukariotic genomes, the present case is to our knowledge the only fully characterised natural example. Inversions occur in prokaryotes as functionally important switching mechanisms [21,22] and in conjuction with deletion rearrangements [10,11] , and are almost invariably associated with flanking inverted repeats, a feature which we failed to find for this haemoglobin recombinant. A functional role for inversion and associated deletion in eukaryotes has, however, been proposed as a mechanism whereby the functional immunoglobulin gene is formed in lymphocytes from linked, but widely seperated gene segments; an overall pattern of rearrangement very similar to that found in the present beta globin gene cluster. Such an inversion-deletion mechanism has been demonstrated with retrovirus constructs containing immunoglobulin gene segments and their signal sequences integrated into the genome of lymphocytes, showing that such inversions within genomic DNA can occur quite readily, and if the model proves correct that this mode of rearrangement may not be so very unusual. Immunoglobulin gene recombination requires specific signal sequences consisting of 7 and 9 nucletides closely conforming to an overall consensus sequence. However aberrant immunoglobulin

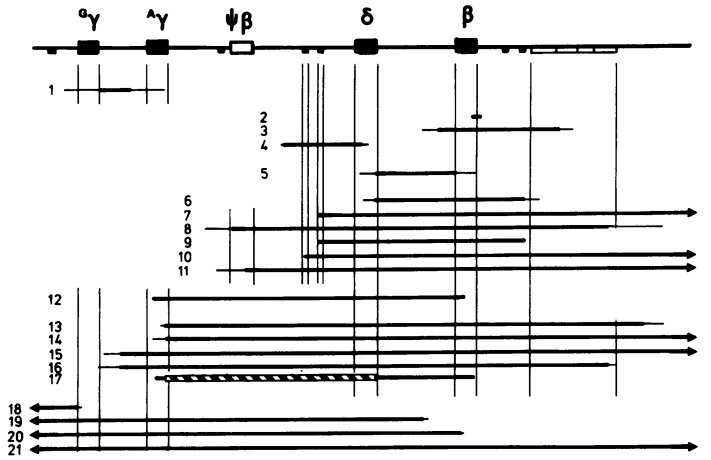
rearrangements are known which appear to involve the recombination of an immunoglobulin gene segment and another, non immunoglobulin coding, region which contains nucleotide sequences resembling the immmunoglobulin recombination signal heptamer [23-25]. These immununoglobulin rearrangements appear to involve an enzyme called endonuclease J [26,27] which recognises the heptamer signal sequence. However the lack of tissue specificity and loose specificity of cleavage, has led to the suggestion that this enzyme may be involved in general recombination, subject to further modulation. From this point of view it is interesting to note that heptamers occur near both breakpoints in the $^{A}\gamma$ gene showing the same degree of conformity to the consensus signal sequence as do actual immunoglobulin signals. However such a match of 6 of 7 to the consensus are relatively common in the region, occurring at a frequency close to that expected of 1/750 and these heptamers do not show the specific patter of rearrangement seen for immunoglobulin gene segments.

The comparison of the two types of rearrangements shows that inversion-deletion rearrangements can occur quite readily, and that there may exist a generalised mechanism for this which could in some unusual event give rise to rearrangements like the one in the β complex, subject to other influences and constraints such as sequence homology, chromatin structure and folding, or pattern of DNA replication.

Expression of foetal globin genes.

This globin gene rearrangement has several implications for the mechanism by which deletions involving the δ and β genes leave the upstream fetal γ genes active in adult life. Thirteen deletions within the β-globin gene cluster that result in enhanced γ gene expression (with respect to that in Hb Lepore) have been mapped, and the question arises as to whether there are any features which distinguish these from the three deletions without elevated γ-gene expression, and further whether the postion or extent of those deletions could account for the different degrees of increased γ chain production which distinguish the δβ-thalassemias and HPFH's (figure 4).

Amongst proposed models, attention has focused on the region 3' to the β gene containing intermediate repeat DNA, which is absent in most rearranged β clusters with elevated γ activity, and present in clusters with quiescent γ genes in the adult [28]. However in the rearrangement described in this paper the region 3' to the β gene seems to be completetly intact, and taken with the Hb Kenya rearrangement these are important exceptions to this rule. One feature, however, common to all clusters with enhanced γ gene expression is



Key	Disorder	Ref.	Key	Disorder	Ref.
1.	γ-Thalassemia	[35]	12.	Hb Kenya	[2]
2.	Indian $(\beta)^o$	[9]	13.	Black $(^A\gamma\delta\beta)^o$	[32]
3.	Dutch $(\beta)^o$	[28]	14.	Chinese $(^A\gamma\delta\beta)^o$	[43]
4.	Greek $(\beta\delta)^o$	[36]	15.	Malasian $(^A\gamma\delta\beta)^o$	[33]
5.	Hb Lepore	[2]	16.	Turkish $(^A\gamma\delta\beta)^o$	[31,37]
6.	Sicilian $(\delta\beta)^o$	[31,37]	17.	Indian $(^A\gamma\delta\beta)^o$	[1]
7.	Spanish $(\delta\beta)^o$	[30,38]	18.	$(^G\gamma^A\gamma\delta\beta)^o$	[44]
8.	Indian HPFH/$(\delta\beta)^o$	[39,40]	19.	Dutch $(^G\gamma^A\gamma\delta\beta)^o$	[45]
9.	Black HPFH	[29]	20.	Anglo-Saxon $(^G\gamma^A\gamma\delta\beta)^o$	[14,46,47]
10.	HPFH I	[14,37,41]	21.	$(^G\gamma^A\gamma\delta\beta)^o$	[48]
11.	HPFH II(Ghanian)	[14,37,42]			

Figure 4. Part of the β cluster showing the positions of Alu and KpnI family repeat sequences (respectively filled and open boxes on the underside of the horizontal) in relation to the majority of characterised deletions affecting the cluster, indicated underneath by solid lines, boldly where deletions of the corresponding region is certain, and less boldly for the regions where the exact end point has not been determined. The vertical lines have been added to clarify the relationship of the deletion end points to genes and repeated sequences.

the removal of sequences within the 5' end of the β gene

Another region of interest is that between the γ and δ genes in the region of the pair of bipolar Alu family repeat sequences. Deletions which leave this region intact result in moderate levels of γ gene expression, whereas with loss of this region the output per γ gene is doubled in the two types of African HPFH. In the present rearrangement this region between the $A\gamma$ and δ genes remains intact yet $G\gamma$ gene expression is as high as in the HPFH cases. Thus if control sequences within this region are of importance, their inversion must render them inactive. Recently, however, this theory has been weakened with the characterisation of two further deletions, one of which

completely removes this region, but does not result in any γ gene enhancement, and a second where there is is great γ enhancement, but in which this region remains intact [29].

Clearly regulation of this gene cluster may involve control processes acting at more that one site in the cluster, each open to interference by various genetic lesions, whether through the loss of specific recognition sequences, disruption of local chromatin structure, or affecting trans-acting factors, and it may be unreasonable to expect a single mechanism to be responsible for increased HbF production in all the aberrently expressed clusters. Whatever the case, the local disruption in this cluster caused by two small deletions and an inversion is phenotypically indistinguishable from four different linear deletions in which the same region and extensive sequences downstream are totally removed [30-33].

REFERENCES

1 Jones,R.W., Old,R.J., Trent,R.J., Clegg,J.B., and Weatherall,D.J. (1981). Nature 291, 39-44.
2 Weatherall,D.J., and Clegg,J.B. (1981). The thalassaemia syndromes. 3rd Edn. Blackwell Scientific Publications, Oxford, England.
3 Amin,A.B., Pandya,N.L., Diwin,P.P., Darbre,P.D., Kattamis,C., Metaxatou-Mavromati,A., White,J.M., Wood,W.G., Clegg,J.B., and Weatherall,D.J. (1979). Br. J. Haematol. 43,537-548.
4 Maniatis,T., Fritsch,E.F., and Sambrook,J. (1982). Molecular cloning. A laboratory manual. Cold Spring Harbor Laboratory. 1982 USA
5 Maxam,A.M., and Gilbert,W. (1980). Methods in Enzymology, 65,499-580.
6 Shen,S., Slightom,J.L.,and Smithies,O. (1981). Cell 26, 191-203.
7 Poncz,M., Schwartz,E., Ballantine,M., and Surrey,S. (1983). J. Molec. Biol. 258, 11599-11609.
8 Staden,R (1984). Nucl. Acids Res. 12,521-538.
9 Spritz,R., and Orkin,S.H. (1982). Nucl. Acids Res. 10, 8025-8028.
10 Botchan,M., Stringer,J., Mitchison,T., and Sambrook,J. (1980). Cell 20,143-152.
11 Ruley,H.E., and Fried,M. (1983) Nature 304, 181-184.
12 Anderson,R.A., Kato,S., and Camerini-Otero,R.D. (1984). Proc. Natn. Acad. Sci. U.S.A. 81, 206-210.
13 Tonegawa,S. (1983). Nature, 302, 575-581.
14 Vanin,E.F., Henthorn,P.S., Kioussis,D., Grosveld,F., and Smithies,O. (1983). Cell 35,701-709.
15 Subrumani,S., and Berg,P. (1983). Molec. Cellul. Biol. 3,1040-1052.
16 Johnson,A.D., Barkan,A., and Mertz,J.E. (1982). Virology, 123,464-469.
17 Nicholls,R.D., Higgs,D.R., Clegg,J.B., and Weatherall,D.J. (1984). Blood, in press
18 Bullock,P. and Botchan,M. (1983). In "Gene Amplification",215-224,(ed. Schimke, R.T.)(Cold Spring Harbour ,NY, 1982).
19 Vitelli,L. and Weinberg,E.S. (1983). Nucl. Acids Res. 11,2135-2153.
20 Wang,S-W, Robins,A.J., d'Andrea,R. and Wells,J.R.E. (1985) Nucl. Acids Res. 13,1369-1387.
21 Davison,A.J., and Wilkie,N.M. (1983) J. Gen. Virol. 64,1-18.
22 Plasterk,R.H.A., Kanaar,R., and van de Putte,P. (1984). Proc. Natn. Acad. Sci. U.S.A. 81,2689-2692.
23 Hochtl,J., and Zachau,H.G. (1983). Nature 302,260-263.
24 Seidman,J.G., and Leder,P. (1980). Nature 286,779-783.
25 Leder,P., Max,E.E., Seidman,J.G., Kwan,S., Scharff,M., and Norman,B. (1981). Cold Spring Harb. Symp. Quant. Biol. 45,859-865.
26 Kataoka,T., Kondo,S., Nishi,M., Kodaira,M., and Honjo,T (1984). Nucl. Acids Res. 12,5995-6010.
27 Desiderio,S. and Balitmore,D. (1984). Nature 308, 860-862.
28 Gilman,J.G., Huisman,T.H.J., and Abels, J. (1984). Br. J. Haematol. 56,339-348.
29 Anagnou, N.P., Papayannopoulou, T., Stamatoyanopoulos, G., and Nienhuis, A.W. (1984). Clinical Research 32,303A.

30 Ottolenghi,S, et al. (1982) Proc. Natn. Acad. Sci. U.S.A. 79, 2347-2351.
31 Fritsch,E.F., Lawn,R.M., and Maniatis,T. (1980). Cell 19, 959-972.
32 Henthorn,P.S., Smithies,O., Nakatsuji,T., Filice,A.E., Gardiner,M.B. Reese, A.L., and Huisman,T.J.H. (1984). in press.z
33 Trent,R.J., Jones,R.W., Clegg,J.B., Weatherall,D.J., Davidson,R., Wood,W.G. (1984) Br. J. Haematol. 57, 279-289.
34 Slightom,J.L., Blechl,A.E., and Smithies,O. (1980). Cell 21, 627-638.
35 Sukumaran,P.K., Nakatsuji,T., Gardiner,M.B., Reese,A.L., Gilman,J.G., and Huisman,T.H.J. (1983). Nucl. Acids Res. 11,4635-4643.
36 Wainscoat,J.S., Thein,S.L., Wood,W.G., Weatherall,D.J., Metazotou-Mavromati,A., Tzotos,E., Kanavakis,E., and Kattamis. (1984) ANYAS In Press
37 Tuan,D., Feingold,E., Newman,M., Weissman,S.M., and Forget,B.G. (1983). Proc. Natn. Acad. Sci. U.S.A. 80,6937-7941.
38 Ottolenghi,S, and Giglioni. (1982). Nature 300, 770-771.
39 Kutlar,A., Gardiner,M.B., Headlee,M.G., Reese,A.L., Cleek,M.P.,Nagle,S., Sukumaran,P.K., and Huisman,T.H.J. (1984). Biochem. Genet. 22, 21-35
40 Wainscoat,J.S., Old,J.M., Wood,W.G., Trent,R.J., and Weatherall, D.J. (1984) Br. J. Haematol. In Press
41 Jagadeeswaran,P., Tuan,D., Forget,B.G., and Weissman,S.M. (1982). Nature 296, 469-470
42 Bernards,R., and Flavell,R.A. (1980). Nucl. Acids Res. 8, 1521-1534.
43 Jones,R.W., Old,J.M., Trent,R.J., Clegg,J.B., and Weatherall, D.J. (1981) Nucl. Acids Res. 9,6813-1825.
44 Piratsu,M., Curtin,P., Kan,Y.W., Gobert-Jones,J.A., Stephens,A.D., and Lehmann,H. (1984). Clinical Res. 32,493A.
45 Van de Ploeg,L.H.T., Konings,A., Oort,M., Roos,D., Bernini,L., and Flavell,R.A. (1980). Nature 283, 637.
46 Orkin,S.H., Goff,S.C., and Nathan,D.G. (1981). J. Invest. Med. 67, 878-884.
47 Kan,Y.W., Forget,B.G.,andNathan,D.G. (1972) New Engl. J. Med. 286,129-134.
48 Fearon,E.R. et al. (1983). Blood 61,1273-1278.